

关于系统聚类法的若干注记

费 荣 昌

(基础课部)

摘 要

本文从各种距离之间的关系来对重心法、类平均法及离差平方和法作一些分析比较。

系统聚类法是目前使用最多的聚类分析方法,类与类之间的距离的定义不同,就产生系统聚类的不同方法,在文献[1][2][3]中介绍了八种方法,有关这些方法的优劣有一些评论,理论上并未彻底解决。本文从各种距离之间的关系来对重心法、类平均法及离差平方和法作一些分析比较。

设类 G_p 包含的样品个数为 n_p , 重心为 $\bar{X}_p$, 离差平方和为 S_p ; 类 G_q 包含的样品个数为 n_q , 重心为 $\bar{X}_q$, 离差平方和为 S_q 。以 $D_{pq}^2(\text{重})$ 、 $D_{pq}^2(\text{平})$ 、 $D_{pq}^2(\text{离})$ 分别表示重心法、类平均法及离差平方和法的类 G_p 与类 G_q 之间的平方距离, 它们的定义为

$$D_{pq}^2(\text{重}) = (\bar{X}_p - \bar{X}_q)' (\bar{X}_p - \bar{X}_q),$$

$$D_{pq}^2(\text{平}) = \frac{1}{n_p n_q} \sum_{\substack{X_i \in G_p \\ X_j \in G_q}} d_{ij}^2,$$

$$D_{pq}^2(\text{离}) = S_r - S_p - S_q,$$

其中 d_{ij} 是样品 X_i 与样品 X_j 之间的距离, S_r 是由 G_p 与 G_q 合并成的类 G_r 的离差平方和。

我们在 d_{ij} 采用欧氏距离下, 先来建立 $D_{pq}^2(\text{重})$ 与 $D_{pq}^2(\text{平})$ 之间的关系。

关系 1
$$D_{pq}^2(\text{平}) = D_{pq}^2(\text{重}) + V_p + V_q,$$

其中

$$V_p = \frac{1}{n_p} \sum_{X_i \in G_p} (X_i - \bar{X}_p)' (X_i - \bar{X}_p), \quad V_q = \frac{1}{n_q} \sum_{X_j \in G_q} (X_j - \bar{X}_q)' (X_j - \bar{X}_q)$$

分别称为 G_p 、 G_q 的方差。

$$\begin{aligned}
 \text{证} \quad D_{pq}^2(\text{平}) &= \frac{1}{n_p n_q} \sum_{X_i \in G_p} \sum_{X_j \in G_q} (X_i - X_j)' (X_i - X_j) \\
 &= \frac{1}{n_p n_q} \sum_{X_i \in G_p} \sum_{X_j \in G_q} (X_i - \bar{X}_p + \bar{X}_p - X_j)' (X_i - \bar{X}_p + \bar{X}_p - X_j) \\
 &= \frac{1}{n_p} \sum_{X_i \in G_p} (X_i - \bar{X}_p)' (X_i - \bar{X}_p) + \frac{1}{n_q} \sum_{X_j \in G_q} (\bar{X}_p - X_j)' (\bar{X}_p - X_j) \\
 &\quad + \frac{2}{n_p n_q} \sum_{X_i \in G_p} (X_i - \bar{X}_p)' \sum_{X_j \in G_q} (\bar{X}_p - X_j),
 \end{aligned}$$

注意最后一项等于零, 即有

$$\begin{aligned}
 D_{pq}^2(\text{平}) &= V_p + \frac{1}{n_q} \sum_{X_j \in G_q} (\bar{X}_p - \bar{X}_q + \bar{X}_q - X_j)' (\bar{X}_p - \bar{X}_q + \bar{X}_q - X_j) \\
 &= V_p + (\bar{X}_p - \bar{X}_q)' (\bar{X}_p - \bar{X}_q) + \frac{1}{n_q} \sum_{X_j \in G_q} (\bar{X}_q - X_j)' (\bar{X}_q - X_j) \\
 &\quad + \frac{2}{n_q} (\bar{X}_p - \bar{X}_q)' \sum_{X_j \in G_q} (\bar{X}_q - X_j) \\
 &= V_p + D_{pq}^2(\text{重}) + V_q
 \end{aligned}$$

$D_{pq}^2(\text{重})$ 仅反映类 G_p 的重心 $\bar{X}_p$ 与类 G_q 的重心 $\bar{X}_q$ 之间的几何距离, 而由关系 1 可知 $D_{pq}^2(\text{平})$ 不仅反映 $\bar{X}_p$ 与 $\bar{X}_q$ 之间的几何距离, 并且还反映类 G_p 和类 G_q 所含样品的离散程度, 也就是说, 采用重心法系统聚类的结果仅取决于聚类过程中所涉及的各类的重心之间的几何距离, 而采用类平均法系统聚类的结果不仅取决于各类的重心之间的几何距离, 还取决于各类所含样品的离散程度。

其次来考虑 $D_{pq}^2(\text{重})$ 与 $D_{pq}^2(\text{离})$ 之间的关系。

$$\text{关系 2} \quad D_{pq}^2(\text{离}) = \frac{n_p n_q}{n_r} D_{pq}^2(\text{重}),$$

其中 $n_r = n_p + n_q$

证 用方差分析的知识来证, 把 G_r 分成 G_p 和 G_q 两类, S_r 是总离差平方和, $S_p + S_q$ 是类内离差平方和, 于是 $S_r - S_p - S_q$ 就是类间离差平方和, 有

$$D_{pq}^2(\text{离}) = S_r - S_p - S_q = n_p (\bar{X}_p - \bar{X}_r)' (\bar{X}_p - \bar{X}_r) + n_q (\bar{X}_q - \bar{X}_r)' (\bar{X}_q - \bar{X}_r)$$

利用

$$\bar{X}_r = \frac{n_p \bar{X}_p + n_q \bar{X}_q}{n_r},$$

得

$$\bar{X}_p - \bar{X}_r = \frac{n_q}{n_r} (\bar{X}_p - \bar{X}_q), \quad \bar{X}_q - \bar{X}_r = -\frac{n_p}{n_r} (\bar{X}_p - \bar{X}_q),$$

代入上式, 即得

$$\begin{aligned} D_{pq}^2(\text{离}) &= n_p \cdot \frac{n_p^2}{n_r^2} (\bar{X}_p - \bar{X}_q)' (\bar{X}_p - \bar{X}_q) + n_q \cdot \frac{n_q^2}{n_r^2} (\bar{X}_p - \bar{X}_q)' (\bar{X}_p - \bar{X}_q) \\ &= \frac{n_p n_q}{n_r} (\bar{X}_p - \bar{X}_q)' (\bar{X}_p - \bar{X}_q) = \frac{n_p n_q}{n_r} D_{pq}^2(\text{重}). \end{aligned}$$

由关系 2 可知, $D_{pq}^2(\text{离})$ 不仅反映 $\bar{X}_p$ 与 $\bar{X}_q$ 之间的几何距离, 并且还反映 G_p 和 G_q 所含样品的个数。

当 n_r 固定时, 由

$$\frac{n_p n_q}{n_r} = \frac{n_r^2 - (n_p - n_q)^2}{4n_r}$$

可知, $\frac{n_p n_q}{n_r}$ 随 $|n_p - n_q|$ 减小而增大, 并且在

$$|n_p - n_q| = \begin{cases} 0, & n_r \text{ 为偶数;} \\ 1, & n_r \text{ 为奇数} \end{cases}$$

时达到最大值。

当 n_p 固定时, 由

$$\frac{n_p n_q}{n_r} = \frac{n_p}{\frac{n_p}{n_q} + 1}$$

可知, $\frac{n_p n_q}{n_r}$ 随 n_q 增大而增大。

从而, 在聚类过程中, $\frac{n_p n_q}{n_r}$ 能使各类包含样品的个数都不过分地多, 起了匀分的作用。

在 [3] 中, 把 $\frac{n_p n_q}{n_r}$ 看作对 $D_{pq}^2(\text{重})$ 的加权, 并指出, 这个权可以推广到类与类之间的其它距离上去。

我们对 $D_{pq}^2(\text{平})$ 加上权 $\frac{n_p n_q}{n_r}$, 记

$$D_{pq}^2(1) = \frac{n_p n_q}{n_r} D_{pq}^2(\text{平}) = \frac{1}{n_r} \sum_{\substack{X_i \in G_p \\ X_j \in G_q}} d_{ij}^2$$

在类平均法的递推公式

$$D_{kr}^2(\text{平}) = \frac{n_p}{n_r} D_{kp}^2(\text{平}) + \frac{n_q}{n_r} D_{kq}^2(\text{平})$$

两边乘以 $\frac{n_k n_r}{n_k + n_r}$, 有

$$\frac{n_k n_r}{n_k + n_r} D^2_{kr}(\text{平}) = \frac{n_k + n_p}{n_k + n_r} \cdot \frac{n_k n_p}{n_k + n_p} D^2_{kp}(\text{平}) + \frac{n_k + n_q}{n_k + n_r} \cdot \frac{n_k n_q}{n_k + n_q} D^2_{kq}(\text{平}),$$

即得递推公式

$$D^2_{kr}(1) = \frac{n_k + n_p}{n_k + n_r} D^2_{kp}(1) + \frac{n_k + n_q}{n_k + n_r} D^2_{kq}(1).$$

它比离差平方和法的递推公式简单。

如果对 $D^2_{pq}(\text{平})$ 加上权 $\frac{1}{n_r}$, 记

$$D^2_{pq}(2) = \frac{1}{n_r} D^2_{pq}(\text{平}) = \frac{1}{n_p n_q n_r} \sum_{\substack{X_i \in G_p \\ X_j \in G_q}} d^2_{ij}.$$

把类平均法的递推公式改写成

$$(n_k + n_r) D^2_{kr}(2) = \frac{n_p}{n_r} \cdot (n_k + n_r) D^2_{kp}(2) + \frac{n_q}{n_r} \cdot (n_k + n_q) D^2_{kq}(2),$$

则得递推公式

$$D^2_{kr}(2) = \frac{n_p}{n_r} \cdot \frac{n_k + n_p}{n_k + n_r} D^2_{kp}(2) + \frac{n_q}{n_r} \cdot \frac{n_k + n_q}{n_k + n_r} D^2_{kq}(2).$$

这里, 权 $\frac{1}{n_r}$ 促进聚成大类, 起着不均分的作用。

同样, 可对 $D^2_{pq}(\text{重})$ 和 $D^2_{pq}(\text{离})$ 加上权 $\frac{1}{n_r}$ 。

最后, 作为关系 2 的应用, 我们给出离差平方和法的递推公式的一个简单证法。在重心法的递推公式

$$D^2_{kr}(\text{重}) = \frac{n_p}{n_r} D^2_{kp}(\text{重}) + \frac{n_q}{n_r} D^2_{kq}(\text{重}) - \frac{n_p}{n_r} \cdot \frac{n_q}{n_r} D^2_{pq}(\text{重})$$

两边使用关系 2, 有

$$\begin{aligned} \frac{n_k + n_r}{n_k n_r} D^2_{kr}(\text{离}) &= \frac{n_p}{n_r} \cdot \frac{n_k + n_p}{n_k n_p} D^2_{kp}(\text{离}) \\ &+ \frac{n_q}{n_r} \cdot \frac{n_k + n_q}{n_k n_q} D^2_{kq}(\text{离}) - \frac{1}{n_r} D^2_{pq}(\text{离}), \end{aligned}$$

即得离差平方和法的递推公式:

$$D^2_{kr}(\text{离}) = \frac{n_k + n_p}{n_k + n_r} D^2_{kp}(\text{离}) + \frac{n_k + n_q}{n_k + n_r} D^2_{kq}(\text{离}) - \frac{n_k}{n_k + n_r} D^2_{pq}(\text{离}).$$

可见, 比 [1] 中的证法简单得多。

本文写作过程中曾和李中奇、吴大伟、吴有炜一起讨论过, 表示感谢。

参 考 文 献

- [1] 张尧庭, 方开泰, 多元统计分析引论, 科学出版社, 1982.
- [2] 方开泰, 聚类分析 (I), 数学的实践与认识, 1, 1978.
- [3] 方开泰, 潘恩沛, 聚类分析, 地质出版社, 1982.

Some Notes on Hierarchical Clustering Methods

Fei Rongchang

ABSTRACT

By making an analysis of different relations of distance, method of gravity center, method of mean value of class and method of sum of squares of deviations are compared in this paper.